ICT Technical Report *ICT-TR-03-2005*

Evaluating Social Causality and Responsibility Models: An Initial Report

Wenji Mao Jonathan Gratch

Abstract Intelligent virtual agents are typically embedded in a social environment and must reason about social cause and effect. Social causal reasoning is qualitatively different from physical causal reasoning that underlies most current intelligent systems. Besides physical causality, the assessments of social cause emphasize epistemic variables including intentions, foreknowledge and perceived coercion. Modeling the process and inferences of social causality can enrich believability and cognitive capabilities of social intelligent agents. In this report, we present a general computational model of social causality and responsibility, and empirical results of a preliminary evaluation of the model in comparison with several other approaches.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluating Social Causality and Responsibility Models: An Initial Report	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**15**	

1. Introduction

Research in intelligent virtual agents has emphasized human-like qualities in the physical manifestation of agents, but such realism is typically skin-deep. Although agents can interact in naturalistic ways with human users and can successfully mimic speech, body language and even, the core reasoning techniques that drive such behaviors have not fundamentally changed. Most intelligent systems incorporate planning and reasoning techniques designed to reason about *physical* causality. Unfortunately, physical causes and effects are simply inadequate for exploiting and explaining social phenomena. In contrast, *social causality*, both in theory and as practiced in everyday folk judgments and in the legal system, emphasizes multiple causal dimensions, incorporates epistemic variables, and distinguishes between cause, responsibility and blame.

Recent approaches to social causality have addressed these differences by extending causal models [Halpern & Pearl, 2001; Chockler & Halpern, 2004], although it is unclear whether a full accounting of social causality will (or even should) result from such extensions. In contrast, we start with social causality theory and consider how this could be formalized in a computational model. This allows intelligent entities to reason about aspects of social causality not addressed by extended causal models and provides a complementary perspective to the enterprise of causal reasoning about social events.

Psychological and philosophical theories identify key variables that mediate determinations of social causality. In these theories, social causality involves not only physical causality, with an emphasis on human agency, but also people's freedom of choice (e.g., coercion [Shaver, 1985] and controllability [Weiner, 1995]), intentions and foreknowledge [Shaver, 1985; Zimmerman, 1988]. Using these variables, social causality makes several distinctions not present in the determinations of physical cause. For example, an actor may physically cause an event, but be absolved of responsibility and blame. Or a person may be held responsible and blameworthy for what she did not physically cause.

Our goal is to model the underlying process and inferences of social causality to enrich the cognitive and social functionality of intelligent agents. Such a model can help an agent to explain the observed social behavior of others, which is crucial for successful interactions among social entities. It can enrich the design components of human-like agents, guide strategies of natural language conversation and model social emotions [Gratch & Marsella, 2004]. To achieve this end, we base our work on the broad variables people use in determining social causality and responsibility. Psychological and philosophical theories largely agree on these basic variables though they differ in terminology. In this report, we adopt the terminology of Shaver [1985]. In Shaver's model, the judgment process proceeds by assessing several key variables: who *caused* the event; Did the actor *foresee* the consequence; Did she *intend* to bring the consequence about; Did she have *choices* or act under *coercion* (e.g., by an authority)?

Though the theory identifies the conceptual variables for social causality and responsibility judgment, in modeling social behavior of intelligent agents, we cannot assume that an agent has privileged access to the mental states of other agents, but rather, an agent can

only make inferences and judgment based on the evidence accessible in the computational system it situates. Current intelligent systems are increasingly sophisticated, usually involving natural language conversation, interactions of multiple agents and a planning module to plan for sequence of actions, with methods that explicitly model beliefs, desires and intentions of agents. All these should play a role in evaluating the conceptual variables underlying social causality and responsibility judgment.

In order to bridge the conceptual descriptions of the variables and the computational realization in application systems, we need to model the inferential mechanism that derives the variable values needed for the judgment from information and context available in practical systems. This report presents a domain-independent computational model of social causality and responsibility. The model infers the key variables from plan knowledge and communication. To assess the veracity of the approach in modeling human social inference, we conduct empirical studies to evaluate and compare the model with several other models of responsibility and blame.

In the rest of the report, we first introduce the judgment process and how the key variables are utilized in the process, and then present the computational model. We finally evaluate the model using empirical data and compare our approach with the related work.

2. Judgment Process and Key Variables

We base our work on the most influential attributional models of Shaver [1985] and Weiner [1995] for social causality and responsibility. Their models suggest that physical causality and coercion identify *who* is responsible for some outcome under evaluation, whereas mental factors, intention and foreseeability, determine *how much* responsibility and blame/credit are assigned.

The evaluations of physical causality and coercion identify the responsibility party. *Physical causality* refers to the connection between actions and the effects they produce. In the absence of external coercion, the actor whose action directly produces the outcome is regarded as responsible. However, in social situations, an agent may cause an outcome because she could not have done otherwise. *Coercion* occurs when some external force, such as a more powerful individual or a socially sanctioned authority, limits an agent's freedom of choice. The presence of coercion can deflect some or all of the responsibility to the coercive force, depending on the perceived degree of coercion.

Intention and forseeability determine the degree of responsibility and blame. *Intention* is generally conceived as the commitment to work towards a certain act or outcome. Most theories view intention as the major determinant of the degree of responsibility. If an agent intends an action to achieve an outcome, then the agent must have the foreknowledge that the action brings about the outcome. *Foreseeability* refers to an agent's foreknowledge about actions and their consequences. The higher the degree of intention, the greater the responsibility assigned. The lower the degree of foreseeability, the less the responsibility assigned.

An agent may intentionally perform an action, but may not intend all the action effects. It is *outcome intent* (i.e., intentional action effect), rather than *act intentionality* (i.e., intentional action) that are key in responsibility judgment [Weiner, 2001]. Similar difference exists in *outcome coercion* (i.e., coerced action effect) and *act coercion* (i.e., coerced action). An agent's intentional action and action effect may succeed or fail. However, as long as it manifests intentions, a *failed attempt* can be blamed or credited almost the same as a successful one [Zimmerman, 1988].

The result of the judgment process is the assignment of certain blame or credit to the responsible agent(s). The intensity of blame or credit is determined by the degree of responsibility as well as the severity or positivity of the outcome. The degree of responsibility is based on the assessed values of attribution variables.

3. The Social Inference Model

We build a computational model of this social judgment process, showing how automated methods for causal and dialogue reasoning can provide a mechanistic explanation of how people arrive at judgments of blame and responsibility. Here we briefly summarize the model. The reader may refer to [Mao & Gratch, 2003a, 2003b, 2004a, 2004b] for details.

3.1 Modular Structure

The judgment of social causality and responsibility is a subjective process. It is from the perspective of a perceiving agent (i.e., the agent who makes the judgment), and based on the perceiver's interpretation of the significance of events. The perceiver uses her own knowledge about the observed agents' behavior to infer certain beliefs (in terms of the key variables). The inferred variable values are then applied to the judgment process to form an overall result.

Two important *sources* of information contribute to the inference of key variables. One source is general beliefs about actions and their effects. The other is observations of the actions performed by the observed agents, including physical and communicative acts (e.g., in a conversational dialogue). The inference process acquires beliefs from communicative events (i.e., dialogue inference) and from the causal information about the observed action execution (i.e., causal inference). To construct a computational model, we need to represent such information and make inferences over it. We also need an algorithm to describe the overall judgment process.

We have designed a modular structure for evaluating social causality and responsibility (i.e., a social inference module), and its interface with other system components. *Figure 1* illustrates the structure of the module. It takes the observed communicative events and executed actions as inputs. *Causal information* and *social information* are also important inputs. Causal information includes an action theory and a plan library (discussed below). Social information specifies social roles and the power relationship of the roles. The *inference* process first applies dialogue inference, and then causal inference. Both make use

of the commonsense heuristics, and derive beliefs about the variable values. The values are then served as inputs of the *algorithm*, which determines responsibility, and assigns certain blame or credit to the responsible agents.

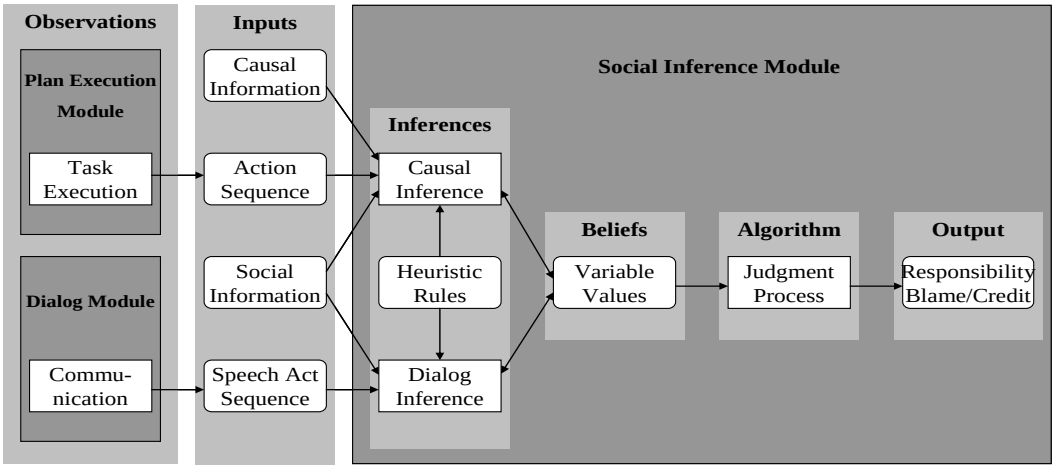

Figure 1 Structure of the Social Inference Module

3.2 Computational Representation

To represent an agent's causal knowledge, we have adopted a hierarchical plan representation used in many intelligent agent systems. This representation provides a concise description of the physical causal relationship between events and world states. It also provides a clear structure for exploring alternative courses of actions and detecting plan interactions.

Actions and Plans

Physical causality is encoded via a hierarchical plan representation. *Actions* consist of a set of propositional preconditions and effects (including conditional effects). Each action step is either a *primitive* action (i.e., an action directly executable by some agent) or an *abstract* action. An abstract action may be decomposed in alternative ways and the effects of an abstract action depend on these alternatives. For example, if there are two alternatives for performing an abstract action, only those effects that occur in each alternative are necessarily the effects of the abstract act. The desirability of action effects is represented by utility values [Blythe, 1999].

A *plan* is a set of actions to achieve certain intended goal(s). As a plan may contain abstract actions (i.e., an abstract plan), each abstract plan indicates a *plan structure* of decomposition. Decomposing the abstract actions into primitive ones in an abstract plan results in a set of primitive plans (i.e., plans with only primitive actions), which is directly executable by agents. In addition, each action in the plan structure is associated with the *performer* (i.e., agents capable of performing the action) and the *authority* (i.e., agent who authorizes the action execution). The performer cannot execute the action until

authorization is given by the authority. This represents the hierarchical organizational structure of social agents.

Communicative Events

Natural language communication is a rich information source for inferring attribution variables. We assume conversations between agents are *grounded* and they conform to Grice's maxims of *Quality* and *Relevance* (i.e., true and relevant information exchange in conversation). We represent communicative events as *speech act* [Austin, 1962] sequence, and analyze the following acts that are typical in negotiation dialogues [Traum *et al*, 2003], *inform*, *order*, *request*, *accept*, *reject*, and *counter-propose*.

3.3 Inferences

The inference of physical causality, coercion, intentions and foreknowledge is informed by dialogue and causal evidence in social interactions. We introduce commonsense heuristics that allow an agent to make inferences based on this evidence.

Agency

A first step in attributing responsibility and blame is to identify which actors' actions contribute to the occurrence of an outcome under evaluation. In a multi-agent plan execution environment, an actor can produce an outcome through the assistance of other agents. These other agents are viewed as indirect agency. Given a specific outcome *p* and the observed action set *S*, the following actions in *S* are relevant to achieving *p*:

- The primitive action *A* that has *p* as its effect.
- The actions that establish a precondition of a relevant action to achieving *p*.
- If *p* or a precondition of a relevant action is enabled by the consequent of a conditional effect, the actions that establish the antecedent of the conditional effect are relevant.

These relevant actions are the possible causes of the outcome *p*. Therefore, their performers are potentially responsible for *p*.

Coercion

An actor could be absolved of responsibility if she was coerced by other forces, but just because an agent applies coercive force does not mean coercion actually occurs. What matters is whether this force truly constrains the actor's freedom to avoid the outcome. Causal inference helps evaluate outcome coercion from evidence of act coercion.

Two concepts are important in understanding coercion. One concept is social *obligation*, created by utterance, role assigned, etc. The other is (*un*)*willingness*. For example, if some authorizing agent commands another agent to perform a certain action, then the latter agent has an obligation to do so. But if the agent is actually willing to, this is a voluntary act rather than a coercive one.

If there is no clear evidence that an agent intends beforehand, and the agent accepts her obligation, there is evidence of coercion. In this inference rule, intend(*x, p, t1*) represents that agent *x* intends that proposition *p* at time *t1*, obligation(*x, p, y, t2*) represents that *x* has an obligation *p* by agent *y* at time *t2*, accept(*x, p, t3*) represents that *x* accepts that *p* at time *t3*, and coerce(*y, x, p, t4*) represents that *y* coerces *x* that *p* at time *t4*.

$\neg(\exists t1)(t1<t3 \wedge \text{intend}(x, p, t1)) \wedge \text{obligation}(x, p, y, t2) \wedge \text{accept}(x, p, t3) \wedge t2<t3<t4 \Rightarrow \text{coerce}(y, x, p, t4)$

In another case, when there is clear evidence of the unwillingness (i.e., intend(*x, p, t1*) is false), there is *strong* evidence of coercion.

Given the action preconditions are initially true, if an agent is coerced to execute a primitive action, the agent is also coerced to achieve all the action effects. If being coerced to execute an abstract action and the action has only one decomposition, then the agent is also coerced to execute the sub-actions and achieve all the sub-action effects. If the coerced action has multiple decompositions, then the agent has options: only the effects appear in all alternatives are unavoidable, and thus these effects are coerced; Since other effects that only appear in some (but not all) alternatives are avoidable, they are not coerced.

If some agents block other action alternatives (by disabling action preconditions), the only alternative left as well as its effects are coerced. These blocking agents are also viewed as coercers. If a conditional effect is coerced and its antecedent is initially true or enabled by other agents, then its consequent is coerced. These other agents are also viewed as coercers. Otherwise, the consequent is not coerced.

Intentions

Intentions play a central role in determining the degree of responsibility and blame assignments. Act and outcome intentions can be inferred from conversation communication between agents. For example, an *order* or a *request* shows the speaker's *intent*. The two speech acts have different implications on the social status between the speaker and the hearer. If an order is successfully issued to a subordinate, it creates a social *obligation* for the subordinate to perform the content of the act. The hearer may *accept*, *reject* or *counter-propose*. Various inferences can be made depending on the response of the hearer and the power relationship between the speaker and the hearer. For example, if the hearer counters the order, and proposes another alternative, it can be inferred that both the speaker and the hearer *know* the *alternatives*. It is also believed that the hearer does not *intend* what is ordered, but *want* the alternative. If the speaker has known the alternatives yet still orders one, infer that the speaker *intends* the chosen action but *not* the alternative. The reader may refer to [Mao & Gratch, 2003b] for the complete rules.

$\text{intend}(s, p, t1) \wedge \neg\text{obligation}(h, p, s, t2) \wedge \text{accept}(h, p, t3) \wedge t1<t3 \wedge t2<t3<t4 \Rightarrow \text{intend}(h, p, t4)$

Outcome intent can also be partially inferred from evidence of act intentionality. For example, if an agent intends an action voluntarily, the agent must intend at least one action effect. If there is only one action effect (significant to the agent), we can exactly infer which effect the agent intends. As plans provide context in evaluating intention, with association to the goals and reasons of an agent's behavior, in the absence of clear evidence from dialogue inference, we employ a general plan-based algorithm to recognize intentions [Mao & Gratch, 2004b].

Foreknowledge

Since foreknowledge refers to an agent's epistemic state, it is mainly derived from dialogue inference. For example, *inform* gives evidence that the conversants know the content of the act. Besides, intention recognition also helps infer an agent's foreknowledge, as intentions entail foreknowledge (*Axiom 4* in [Mao & Gratch, 2004a]).

3.4 Algorithm

The judgment process begins with some specific outcome that is under evaluation, and the judgment result is based on the inferences of variable values introduced above. The acquired values for agency and coercion contribute to the evaluation of responsible agents. We have developed an algorithm for tracing the responsible agents [Mao & Gratch, 2003b]. The algorithm starts with the primitive action that directly causes the evaluated outcome and works up the plan hierarchy. During each loop, it applies inference rules and intention recognition method to reason about attribution variables. If outcome coercion is true, the algorithm proceeds until reaching the root of the plan hierarchy. In the meantime, the application of inference rules and intention recognition algorithm acquires beliefs for foreknowledge and act/outcome intentions, which determine the intensity of responsibility and blame/credit.

4. Evaluation and Comparison

Our *claim* of evaluation is that this model will better predict human judgments of responsibility and blame than other potential approaches. Here, we report the results of an experiment comparing our model and three computational alternatives to human data.

4.1 Alternative Models

It is not uncommon to use physical causality as a substitute for modeling social causality. This was the approach used, for example, in the *MRE* team training system [Rickel *et al*, 2002]. A *simple cause model* always assigns responsibility and blame to the actor whose action directly produces the outcome.

Instead of always picking up the actor, a slightly more sophisticated model can choose the highest authority (if there is one) as the responsible and blameworthy agent. We call such model *simple authority model*.

Chockler and Halpern [2004] propose a structural-model approach to responsibility and blame (abbreviated to *C&H model* below). They give a definition of responsibility, which extends the definition of causality introduced by Halpern and Pearl [2001]. For example, if a person wins an election 11-0, then each voter who votes for her is a cause for the victory, but each voter is less responsible for the victory than if she had won 6-5. Based on this notion of responsibility, they then defined the degree of blame, using the expected degree of responsibility weighed by the epistemic state of an agent.

4.2 Method

Our model argues that people will view blame differently based on their perception of key variables such as intentions and coercion. Thus a good test is to see how the models perform when such variables are systematically manipulated. We compare attributions of blame by the four models with human judgments using four variants of the "firing squad" scenario in [Chockler & Halpern, 2004].

Scenario 1 is the original example: There is a ten-man firing squad. Only one marksman has live bullets in his rifle; the rest have blanks. The marksmen do not know who has the live bullets. They shoot at the prisoner and the death occurs. *Scenario 2* extends the example to include an authority - the commander, who orders the squad to shoot. *Scenario 3* further extends the example by presenting a negotiation dialogue between the commander and the marksmen. The marksmen first reject the commander's order. The commander insists and orders again. Finally the squad accepts the order and shoot. In *Scenario 4*, the commander still orders. However, each marksman has freedom to choose either using blanks or live bullets before shooting.

In each scenario, we query 27 subjects (mostly university staffs including graduates, with ages ranging from 20 to 45 and evenly distributed genders) to assess their judgments of responsibility, blame and coercion.

4.3 Results

Figure 2 shows proportions of the subjects that attribute blame and responsibility to different categories of agents in the scenarios, and corresponding confidence intervals for large population (α=0.05) [Rice, 1994]. For example, in *scenario 1*, 3 subjects blame the marksman with live bullets in his rifle, 19 blames all the marksmen and the rest do not blame any of them. The analysis of the sample data and their confidence intervals show that a small percentage of the population will blame the marksman with live bullets, a significant majority will blame all the marksmen, and a small percentage won't blame any, with 0.95 confidence.

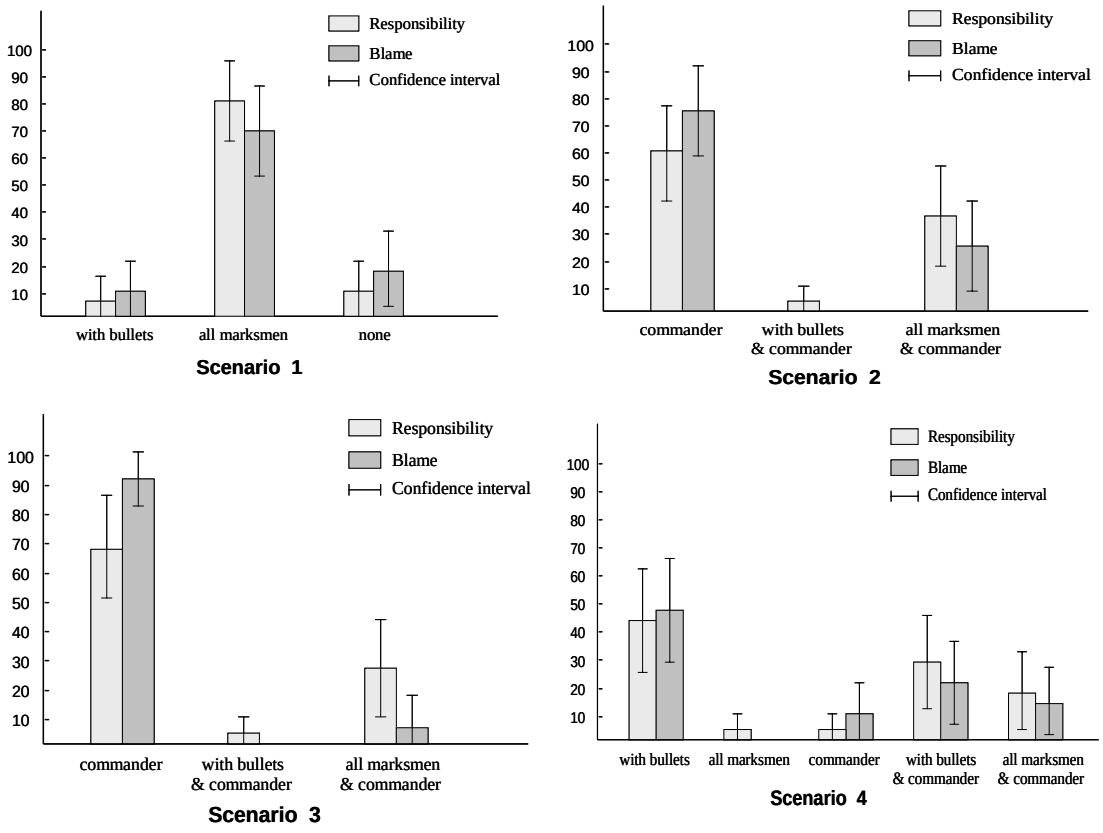

Figure 2 Proportions of Population Agreement on Responsibility/Blame in Sample Scenarios

BLAME	Simple Cause Model		Simple Authority Model		C&H Model		Social Inference Model		Human Majority Agreement
	Results	Match	Results	Match	Results	Match	Results	Match	
S 1	with bullets	no	N/A	N/A	all marksmen	yes	all marksmen	yes	**all marksmen**
S 2	with bullets	no	commander	yes	all marksmen& commander	no	commander	yes	**commander**
S 3	with bullets	no	commander	yes	all marksmen& commander	no	commander	yes	**commander**
S 4	with bullets	yes (partial)	commander	no	N/A	N/A	with bullets	yes (partial)	**with bullets/ w. bullets & commander**

Table 1 Comparison of Results by Different Models with Human Data

Table 1 shows the results on blame generated by different models. All the results are compared with the dominant proportions (i.e. majority) of people's agreement (though in *Scenario 4*, there is an overlap between two categories. That's why we note our model as

a partial match). The simple cause model always chooses physical causality. It only partially matches the human agreement in *Scenario 4*, but is inconsistent with the data in *Scenarios 1-3*. Simple authority model always picks up the highest authority. It matches the human data in *Scenario 2 and 3*, but is inconsistent with the data in other scenarios. In general, simple models are insensitive to the changing situation specified in each scenario.

C&H model matches human judgments only in *Scenario 1*. In the remaining scenarios, its results are incompatible with the data. Like other work in causality research, the underlying causal reasoning in C&H model is based on *philosophical* principles (i.e., counterfactual dependencies). Their extended definition of responsibility accounts better for multiple causes and the extent to which each cause contributes to the occurrence of an outcome. However, the results show that their blame model does not match human data well. These empirical findings generally support our hypothesis.

In the next section, we discuss how our model appraises each scenario and compare our approach with C&H model.

4.4 Comparison and Discussions

Scenario 1

Actions and plans are explicitly represented in our approach. In *Scenario 1*, each marksman performs a primitive action, *shooting*. The action has a conditional effect, with the antecedent *live bullets* and the consequent *death*. All marksmen's shooting actions constitute a team plan *squad firing*, with outcome *death*. The team plan is observed executed, and plan outcome occurs. Applying our intention recognition algorithm[1] [Mao & Gratch, 2004b], the marksmen are believed to intend the actions and the only outcome. The marksman with the bullets is the sole cause of the death. This marksman intends the outcome, and thus deserves high degree of responsibility and blame. As other marksmen with blanks also intend the actions and the outcome, and shooting actions are observed executed but the antecedent of the conditional effect is false, their failed attempt can be detected. Therefore, other marksmen are also blameworthy for their attempt (recall that an agent can be blamed/credited for a successfully produced outcome as well as for an unsuccessful attempt).

C&H model judges responsibility according to the actual cause of the event. As the marksman with the bullets is the only cause of the death, this marksman has degree of responsibility 1 for the death and others have degree of responsibility 0. This result is inconsistent with human data. In determining blame, C&H model draws the same conclusion as ours, but their approach is different. They consider each marksman's epistemic state before action performance (corresponding to foreknowledge). There are 10 situations possible, depending on who has the bullets. Each marksman is responsible for one

[1] Note that our intention recognition algorithm is generally applied to a plan library with multiple plans and sequences of actions, which is typical in intelligent agent applications. In this oversimplified example, intention recognition becomes trivial.

situation with degree of responsibility 1. Given that each situation is equally likely to happen (1/10 possibility), each marksman has degree of blame 1/10.

As there is no notion of intention in their model, C&H model uses foreknowledge as the only determinant for blame assignment. This is fine when there is no foreknowledge, as no foreknowledge entails no intention. However, when there is foreknowledge, the blame assigned is high, even if there might be no intentions in the case. For example, if a marksman fires the gun by mistake, without any intention of shooting or attempting the death, in C&H model, still he will be blamed just the same as those who intend.

Scenarios 2 & 3

In our model, we take different forms of social interactions into account. The inference process reasons about the beliefs from both causal and dialogue evidence. *Figure 3* illustrates the team plan of the squad in *Scenarios 2* and *3*, where a commander acts as an authority of the squad (*AND* denotes that the action has only one decomposition).

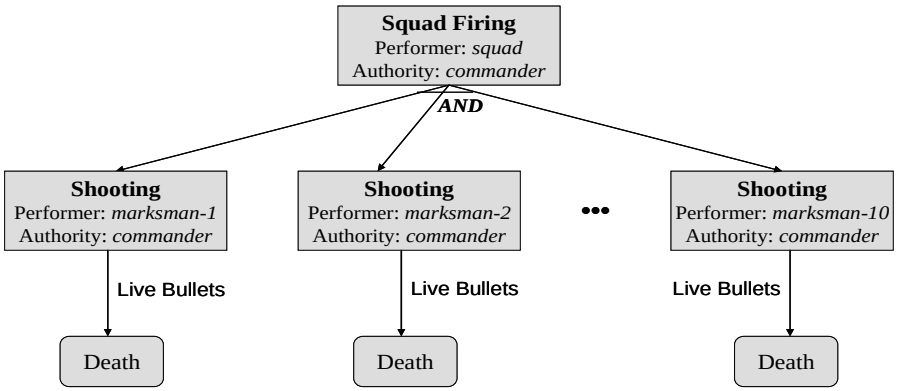

Figure 3 Team Plan of the Squad in Scenarios 2 and 3

The intermediate inference results for *Scenario 2* are given below (*cmd*, *sqd* and *mkn* stand for the commander, the squad and the marksman, respectively. Beliefs are ordered by time).

(1) intend(*cmd*, *do*(*sqd*, *firing*))	(Act *order*)
(2) obligation(*sqd*, *firing*, *cmd*)	(Act *order*)
(3) intend(*cmd*, *death*)	(Rule for *intention* & Result *1*)
(4) coerce(*cmd*, *sqd*, *firing*)	(Act *accept* & Result *2*)
(5) coerce(*cmd*, *mkn*, *shooting*)	(Rule for *coercion*)
(6) coerce(*cmd*, *mkn*, *death*)	(Rule for *coercion*)

So in *Scenario 2*, the marksmen cause/attempt the death due to coercion. The commander is responsible for the death. As the commander intends the outcome, the commander is to blame with high intensity.

Scenario 3 includes a sequence of negotiation acts. The above beliefs 4-6 thus change to the following:

(4) ¬intend(*sqd*, *firing*) (Act *reject* and Result *1*)
(5) coerce(*cmd*, *sqd*, *firing*) (Act *accept* and Results *2* & *4*)
(6) coerce(*cmd*, *mkn*, *shooting*) (Rule for *coercion*)
(7) coerce(*cmd*, *mkn*, *death*) (Rule for *coercion*)

Clearly the marksmen do not intend firing. *Scenario 3* shows strong coercion. This is also reflected in the data. More proportions of people regard the commander as responsible and blameworthy in *Scenario 3* than in *Scenario 2*.

C&H model represents all the relevant events in the scenarios as random variables. So if we want to model the communicative acts in *Scenarios 2* and *3*, each act would be a separate variable in their model. This is problematic when conversational dialogue is involved in a scenario. As the approach uses the structural equations to represent the relationships between variables, and each equation in the model must be deterministic, it is difficult to come up with such equations for a dialogue sequence. For example, if we want to model communicative acts in *Scenario 3*, we will have to give deterministic relationship between them (e.g., if the commander orders, the squad will accept). Such strict equations simply do not exist in a natural conversation. If we ignore some communicative acts in between, important information conveyed by these acts will be lost.

Assume marksman-1 is the one with the live bullets. Using C&H approach, the outcome is counterfactually depends on marksman-1's shooting, so marksman-1's shooting is an actual cause of the death. Similarly, the commander's order is also an actual cause of the death. Based on the responsibility definition in C&H model, both the commander and marksman-1 are responsible for the death, and each has degree of responsibility 1. This result is inconsistent with human data. In assigning blame, there are ten situations altogether, and in each situation, the commander has expected responsibility 1, so the commander is to blame with degree 1. The marksmen each has degree of blame 1/10. Thus C&H model appraises that the commander and all marksmen are blameworthy for the outcome. As their model for responsibility and blame is the extension of counterfactual causal reasoning, which has been criticized as far too permissive [Hopkins & Pearl, 2003], the same problem is also reflected in their model of responsibility and blame.

Scenario 4

Different from the previous scenarios, in *Scenario 4*, the bullets are not initially set before the scenario starts. The marksmen can choose to use either bullets or blanks before shooting. Firing is still the joint action of the squad, but there is no team plan or common goal for the squad. As the commander orders the joint action, act coercion is true. However, based on the rules of inferring outcome coercion from act coercion, the marksmen are not coerced the outcome. So in this case, the commander is not responsible for the outcome, but rather, the marksmen who choose to use bullets and cause the death are responsible and blameworthy. *Figure 2* shows that in *Scenario 4*, people's judgments somehow diffuse. There is an overlap between blaming the marksmen with bullets and blaming both the commander and the marksmen with bullets. Nonetheless, the category our model falls into is clearly better than the rest three.

C&H model requires all the structural equations to be deterministic. In essence, their model could not handle alternative courses of actions, which inherently have nondeterministic property. One way to compensate for this is to push the nondeterminism into the setting of the context. For example, in *Scenario 4*, they could build a causal model to let the context determine whether the bullets are live or blank for each marksman, and then have a probability distribution over contexts. After that, they can compute the probability of an actual cause. However, since these contexts are only background variables, their probabilities could not actually impact the reasoning process per se.

5. Summary

Intelligent virtual agents are typically embedded in a social environment and must reason about social cause and effect. Social causal reasoning is qualitatively different from physical causal reasoning that underlies most current intelligent systems. In this report, we review a general computational model of social causality and responsibility. Our approach bases on the broad features people use in behavior judgment, including physical cause, intentions, foreknowledge and coercion. We present how our model reasons about beliefs about attribution variables for the judgment process, and empirically evaluate and compare the model with several other approaches.

The initial results show that our model better predicts human judgment of blame and responsibility than other potential approaches. Our future work needs to further refine the model and conduct more experiments to systematically evaluate the capabilities of the model.

Acknowledgements

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). We would like to thank Joseph Halpern and Andrew Gordon for the helpful discussions. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

J. Austin. *How to Do Things with Words*. Harvard University Press, 1962.

J. Blythe. Decision-Theoretic Planning. *AI Magazine*, 20(2):37-54, 1999.

H. Chockler and J. Y. Halpern. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22:93-115, 2004.

J. Gratch and S. Marsella. A Domain-Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4):269-306, 2004.

J. Y. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach – Part I: Causes. *Proceedings of the Seventeenth Conference in Uncertainty in Artificial Intelligence*, 2001.

M. Hopkins and J. Pearl. Clarifying the Usage of Structural Models for Commonsense Causal Reasoning. *Proceedings of AAAI Spring Symposium on Logic Formulizations of Commonsense Reasoning*, 2003.

W. Mao and J. Gratch. The Social Credit Assignment Problem. *Proceedings of the Fourth International Working Conference on Intelligent Virtual Agents*, 2003a.

W. Mao and J. Gratch. The Social Credit Assignment Problem (Extended Version). *ICT Technical Report ICT-TR-02-2003*, 2003b.

W. Mao and J. Gratch. Social Judgment in Multiagent Interactions. *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, 2004a.

W. Mao and J. Gratch. A Utility-Based Approach to Intention Recognition. *AAMAS 2004 Workshop on Agent Tracking: Modeling Other Agents from Observations*, 2004b.

J. A. Rice. *Mathematical Statistics and Data Analysis (Second Edition)*. Duxbury Press, 1994.

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum and W. Swartout. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems*, 17(4):32-38, 2002.

K. G. Shaver. *The Attribution of Blame: Causality, Responsibility and Blameworthiness*. Springer-Verlag, 1985.

D. Traum, J. Rickel, J. Gratch and S. Marsella. Negotiation over Tasks in Hybrid Human-Agent Teams for Simulation-Based Training. *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, 2003.

B. Weiner. *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press, 1995.

B. Weiner. Responsibility for Social Transgressions: An Attributional Analysis. In: B. F. Malle, L. J. Moses and D. A. Baldwin (Ed.). *Intentions and Intentionality: Foundations of Social Cognition*, pp. 331-344. The MIT Press, 2001.

M. J. Zimmerman. *An Essay on Moral Responsibility*. Rowman & Littlefield, 1988.